Charles Darwin University

Final Examination

Family Name						
Given Name/s						
Student Number						
Teaching Period	Semester 1, 2018					

PRT562 – Data Analytics and Visualisation	DURATION	
	Reading Time:	10 minutes
	Writing Time:	180 minutes

INSTRUCTIONS TO CANDIDATES

- The examination has **FIVE** questions. Please answer **ALL** questions.
- The total marks of this examination are **50** marks.

EXAM CONDITIONS

You may begin writing from the commencement of the examination session. The reading time indicated above is provided as a guide only.

This is a CLOSED BOOK examination

Any non-programmable calculator is permitted

No handwritten notes are permitted

No dictionaries are permitted

ADDITIONAL AUTHORISED MATERIALS	EXAMINATION MATERIALS TO BE SUPPLIED
No additional printed material is permitted	1 x 16 Page Book 1 x Scrap Paper

THIS EXAMINATION IS PRINTED DOUBLE-SIDED.

THIS PAGE HAS BEEN INTENTIONALLY LEFT BLANK.

Question 1: R Programming [5 marks]

Q1.1 What is the main difference between an array and a matrix in R? [1 mark]
Answer:

Q1.2 How many different types of data object in R? What are they? [1 mark]
Answer:

Q1.3 R statement for creating vectors, called v12 and v13, that contain the following first names and surnames
(Mary, Sam, Beth, George, Helen, Nick, Tracy, David, Jill, Fred)
(Stern, Trill, Matthews, Cray, Beal, Simpson, Deal, Hunter, Wetherby, Sims) [1 mark]
Answer:

Q1.4 In each case below, write down the response (if any) that you would see in the R console window, if the given commands were typed into the console. [2 marks]

a. *> x <- c(1, 1, 2, 3, 5, 8, 13)*

> x[x > 4]

b. *> x <- c(1, 1, 2, 3, NA, 8, 13)*

> x[x == NA] <- 5

> x

Answer:

Question 2 Associate Rule Mining [15 marks]

Q2.1 How are rules generated by APRIORI association rule mining algorithm? How are frequent itemsets used when creating rules? [3 marks]

Answer:

Q2.2 Assume the APRIORI algorithm identified the following seven 4-item sets that satisfy a user given support threshold: acde, acdf, adfg, bcde, bcdf, bcef, cdef. What initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [3 marks]

Answer:

Q2.3 Assume we have an association rule: **if Drink_Tea and Drink_Coffee then Smoke** that has a lift of 2. What does say about the relationship between smoking, and drinking coffee, and drinking tea? Moreover, the support of the above association rule is 1%. What does this mean? [3 marks]

Answer:

Q2.4 Consider the market basket data in the Table as below, [6 marks]

Transaction ID	Items Bought
001	Milk,Beer,Diapers
002	Bread, Butter, Milk
003	Milk, Diapers, Cookies
004	Bread, Butter, Cookies
005	Beer, Cookies, Diapers
006	Milk, Diapers, Bread, Butter
007	Bread, Butter, Diapers
008	Beer, Diapers
009	Milk, Diapers, Bread, Butter
010	Beer, Cookies

a) What is the maximum number of association rules that can be extracted from this data (including rules with zero support)? What is the maximum size (k) of frequent k-itemsets in this data, assuming minsup > 0. [3 marks]

Answer:

b) With same data shown in table above, [3 marks]

- Find an itemset with 2 or more items that has the largest support.
- Find a pair of items a and b, such that the rules a → b and b → a have the same confidence

Answer:

Question 3 Clustering [10 Marks]

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch please show:

1. The new clusters (i.e. the examples belonging to each cluster)
2. The centers of the new clusters
3. Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
4. How many more iterations are needed to converge? Draw the result for each epoch.

Answer:

Question 4 Classification [10 Marks]

Consider the training examples given in Table as below for decision tree binary classification problem.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a) The entropy is given by: [2 marks]

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

where c is the number of classes and p(j|t) is the relative frequency of class j at node t. What is the entropy of this collection of training examples with respect to the positive class?

Answer:

b) What are the information gains of $a1$ and $a2$ relative to these training examples? [2 marks]

Answer:

c) For a3, compute the information gain for every possible split. [3 marks]

Answer:

d) What is the best split (among $a1$, $a2$ and $a3$) according to the information gain and why? [3marks]

Answer:

Question 5 Uncertainty and Reasoning [10 Marks]

Let B stand for "has breast cancer" and M stand for "mammography test is positive." A research study has produced the following three observations.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayes Rule

- The prior probability of having breast cancer is 0.01.
- The probability of testing positive when you have breast cancer is 90%.
- The probability of testing negative when you do not have breast cancer is 89.9%.

a) What is the prior probability of having a positive mammography test? [3 marks]

b) If a patient has a positive mammography test, what is the probability that she has breast cancer? That is, compute P(B|M). [3 marks]

c) If a patient gets a negative mammography test, what is the probability that she has breast cancer? That is, compute P(B|~M). [4 marks]

Answer:

---End of Examination Paper---